ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

# M.Sc. in Data Science

**Course**: Probability and Statistics for Data Analysis

**Semester**: Fall 2018

**Instructor**: Ioannis Vrontos (`vrontos@aueb.gr`)

**Grader**: Konstantinos Bourazas (`kbourazas@aueb.gr`)

# Assignment 1

### Deadline: 15 January 2018

**Note**: Use R in this assignment and submit your .R code that was used to answer the questions, along with a small report where you will present plots and results for each question of this assignment.

**1.** In the spreadsheet named "Data 1" of the file "Assignment_3_Data.xlsx" (available on e-class assignments site) you will find the recorded variables Y, X1, X2, X3 (continuous) and W (categorical with three levels) on 150 cases. Using these data answer the following questions:

**(a)** Run the parametric one-way ANOVA of each of the continuous variables (Y, X1, X2, X3) on the categorical variable (W). Specifically,

**(i)** provide a graphical representation of each of the continuous versus the categorical variable

**(ii)** provide the ANOVA output

**(iii)** check the assumptions and provide alternatives when the assumptions are violated.

**(b)** Provide a scatter-plot matrix of Y, X1, X2, X3, annotating the different levels of W in each plot using a different color.

**(c)** Run the regression model of Y on all the remaining variables (X1, X2, X3, W), including the non-additive terms (i.e. interactions of the continuous predictors with the categorical).

**(d)** Examine the regression assumptions and provide alternatives if any of them fails.

**(e)** Use the "stepwise regression" and the "all subset" approach to examine whether you can reduce the dimension of the model.

**(f)** Using the model found in **(e)**, provide a point estimate and a 95% confidence interval for the prediction of Y when: (X1, X2, X3, W) = (3.1, 3.75, 1.2, A)

**2.** In the spreadsheet named "Data 2" of the file "Assignment 3 data.xlsx" (available on e-class assignments site) you will find the recorded variables Y (continuous) and W, Z (categorical with two levels each) on 84 cases. Using these data answer the following questions:

**(a)** Provide a plot of Y versus the W and Z.

**(b)** Provide the interaction plot of Y versus W and Z.

**(c)** Run the parametric two-way ANOVA of Y on the categorical variables W and Z (including the interaction term) . Provide the fit, examine the assumptions and comment on the significance of the terms.